White Paper Report

Report ID: 106148

Application Number: HD5156012

Project Director: Natalie Houston (nhouston@uh.edu)

Institution: University of Houston

Reporting Period: 5/1/2012-12/31/2013

Report Due: 3/31/2014

Date Submitted: 3/31/2014

Narrative Description

1. Project Activities

This report describes our progress on The Visual Page, a project supported by a Level II Start-Up Grant (HD-51560-12). The goal of this project was to develop a proof-of-concept software application to identify and analyze visual features in digitized books of Victorian poetry.

This project responds to the increasing availability of digitized cultural heritage materials, which increases researchers' access in several ways: digital surrogates can provide access to rare historical documents located in physical archives that may be few in number or geographically disparate; digital surrogates can allow researchers to compare historical documents not available in the same physical location; and digitization can provide the possibility of compiling large data sets for computational analysis.

Central to the theoretical and practical goals of our project is the recognition of the cultural information conveyed by the visual features of printed texts: how page layout, typography, and the distribution of white space visually convey information about the genre, form, purpose, and audience of a text. As Jerome McGann suggests, "A page of printed or scripted text should thus be understood as a certain kind of graphic interface."[1] Experienced readers recognize and interpret the graphic codes of printed pages very rapidly, and those codes influence their reading of the pages' linguistic content. For example, readers can easily distinguish in a book's title page, table of contents, and index from its main content pages. The visual codes of printed books are part of the larger history of the book as a material, cultural object. Because humanities researchers recognize the value of the visual information embedded in the page, most scholarly digital editions today include page images of texts as well as plain text transcriptions of their linguistic content. Our project seeks to make that visual information available for large scale computational analysis, which is already possible for the linguistic content of digitized texts.

We selected books of poetry as our data set for this project because nineteenth-century printing conventions make poetry visually distinct on the page. Some of these conventions include: printing each line of a poem as a separate line of text; grouping lines of poems into stanzas which are separated by white space; printing poetry with generous white space around the text; and indenting lines to indicate the poem's rhyme pattern. These printing conventions visually represent the poem's linguistic and formal or literary features. The human eye can

[1] Jerome McGann, *Radiant Textuality: Literature After the World Wide Web* (New York: Palgrave Macmillan, 2001): 199.

easily distinguish a poem from prose in most nineteenth-century printed texts, and can even recognize particular poetic forms, such as the sonnet. Our project makes that visual information available for computational exploration and analysis.

We were successful in meeting the core goals outlined in our initial proposal: we developed a prototype tool that adapted elements of the open source OCR engine Tesseract to recognize visual elements within document page elements and extracted quantifiable measures of those visual features. Initial analysis of those quantified measures confirmed their relationship to the printed pages and suggested how such measures could be used to explore and analyze large document collections.

We were also successful in communicating our work in progress to several different scholarly communities. These presentations were collaboratively written and jointly presented, which allowed us to model the interdisciplinary collaboration between a literature scholar and a computer scientist that lies at the heart of this project.

2. Accomplishments

2.1 Data Preparation

We data mined bibliographic records using the Online Computer Library Center (OCLC) research API to identify single-author books of poetry published in London between 1860-1880. This large data set was refined to exclude reprint editions of earlier works, so that it reflects new publications by Victorian-era poets. This data set was further limited to items readily available in digitized form from Internet Archive, Google Books, or the HathiTrust Digital Library.

We developed a data preparation process using the General Public License (GPL) Ghostscript tool to extract and convert page images from the PDF files into individual PNG image files. We also developed a file naming and tracking system to manage the data. We extracted book-level bibliographic metadata from the WorldCat cataloging records for use in this project, including author name, book title, place of publication, publisher name, and year of publication.

We then selected a stratified random sample of 75 books published between 1860-1875 to work with during software development.

2.2 Feature Extraction

From a technical perspective, the fundamental task of making the visual structure of documents available for systematic computational analysis involves translating the specific elements of

interest to scholars into features that can be recognized from document images and represented in a structured data format. These features, for example, might include such as line width, height and spacing, margins, or the presence or absence of features such as running heads.

Just as it is necessary to name concepts in order for people to discuss them, feature extraction provides a vocabulary for subsequent algorithmic processing. Once a set of features have been identified as relevant, our application then uses image analysis algorithms to identify and extract metrics for those features (see Figure 1). We can then represent each document as a set of values for these features in much the same way as a bag of words approach is used to represent the linguistic content of a document for use in search engines or topic models. The resulting feature vectors can then be used for subsequent analysis such as to power data visualization, as input for document clustering and pattern recognition software or, as we have done, for custom analysis using a statistical processing language like R.

The process of feature extraction does not begin with a computer but with a scholar. Unlike textual analysis, in which the question of what constitutes a word is straightforward in most instances, the interesting visual elements of meaning are not readily available a document are not distinct, clear cut features. Instead, a scholar first defines a set of visual features according to a theory of the text and a set of research questions. Some of the initial research questions about the initial data set included:

- What patterns in the visual appearance of printed poetry could be seen in a set of documents?
- How consistent or divergent were these features across the pages of individual books or across numerous books in the data set?
- Which books or pages could be considered typical or distinctive, based on selected visual features?

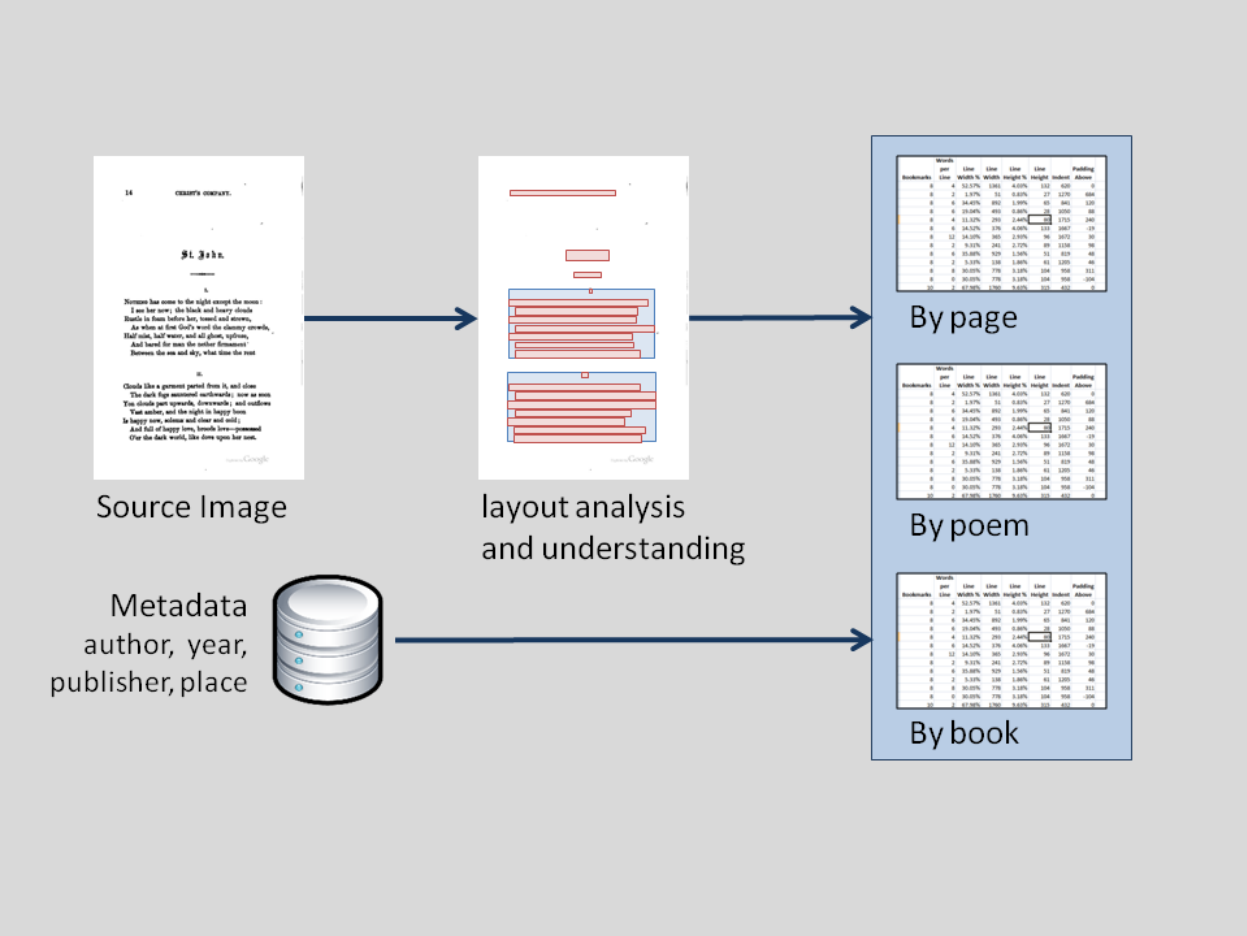


Figure 1

Like other document analysis tasks, identifying the unit of analysis is a key design decision. Four main options stand out:

- **Book:** Books are frequently designed as a cohesive unit with similar visual features used throughout. For understanding issues centered on the technology, economics and sociology of book production, characterizing book-level features such as the presence of running heads, the placement of page numbers and types of illustrations is key.
- **Individual Pages:** Pages are a natural division and the focus of most document image analysis systems.
- **Page Openings:** Since readers typically encounter books as two facing pages, this is the traditional unit for visual design. Page openings are frequently considered the primary unit for representing texts digitally.
- **Natural Divisions:** With respect to book of poetry, the features found within an individual poem, group of poems or sub-unit of a poem such as a stanza, comprise the most important logical divisions for analysis.

During the Start-Up grant period, we focused our efforts on page-based analysis and extracted features related to line-height and width[2] (including both mean and standard deviation), margins, variability of indentation within a page, spacing between lines, and the ratio of foreground pixels to background (effectively the density of ink on the page). In addition to these values, we also recorded the raw data per-line data used to generate these measures.

We used the open source Tesseract OCR engine to perform the low-level image analysis[3]. We created Java-based wrappers that interface directly with this internal component of Tesseract and then used those results to extract the visual features that were relevant for our needs. The results were adequate for our immediate needs and our initial investigation suggested that Tesseract can be used to detect more advanced typographical features such as typeface family and size as well as typeface attributes such as italics, although we did not evaluate the accuracy of those results. Preliminary findings indicated that additional training of Tesseract may be needed. For example, while we intentionally omitted pages with illustrations, some initial tests found that illustrations were frequently misclassified as text regions.

Observations and conclusions

The goal of our work in feature extraction was to determine if the visual features of interest to scholars can be formally described, reliably extracted from document images, and analyzed at scale. We envision eventually creating software tools that allow individual humanities scholars to define a set of visual features that are of interest, create a workflow for extracting those features using pre-existing software modules, and execute that workflow on a set of documents numbering in the thousands to hundreds-of-thousands. Our results were promising, but mixed.

Existing algorithms for low-level image analysis are quite mature and can return a surprising level of detail. Some features, such as the width of a line of text, are directly available from the recognized bounding box of a line of text. Others, such as indentation, can be inferred with post-processing from the structures recognized by text analysis tools, but require careful planning to determine how to measure the feature. For example, we chose to measure the variability of the left-hand bounding box for text lines. More problematic, however, is that existing software typically integrates these layout analysis algorithms into monolithic applications in ways that are not easy to reuse for new purposes. Creating a framework that will allow for the contribution of discrete parts of the image analysis workflow and the

[2] Line width in the purely graphic sense is the visual indication of what is sometimes also termed the (linguistic) length of the line.

[3] As a precursor to character recognition, Tesseract incorporates a state-of-tbe-art image layout analysis system. For a description of this approach, see R. Smith, “Hybrid page layout analysis via tab-stop detection,” In Document Analysis and Recognition, ICDAR'09, 2009, pp. 241-245.

recombination of those components is an area that we have prioritized for future work.

In developing our prototype application, it became obvious that, while page-level visual features do provide value, we must move beyond the page. Understanding the visual rhetoric of a particular poem or a particular book goes beyond merely recognizing the layout of text into regions on a page and involves understanding the semantic structure of the page in domain specific terms: running heads, page numbers, stanzas, titles, epigraphs, paragraphs and more. This necessarily requires prior knowledge of the domain (poetry, prose, or journal articles, for instance). While there is a significant body of literature dedicated to document image understanding, more research is needed into how to apply existing techniques in this area. Moreover, in contrast to DIA, we are not aware of mature, open source systems that have been developed to support image understanding tasks.

The final question is how to make these tools available to a broad audience of humanities scholars. In the course of this initial work, we have gained more insight into the parameters of this multi-faceted challenge. Image analysis of large collections requires vastly more computing power than does text analysis. We will need to design shared infrastructure that does not require data to be delivered over the Internet one image at a time. A second issue is that of training. To pose new questions, scholars need to understand the core technology well enough both to see the possible scope of questions that they can ask and to understand how to properly formulate those questions. Achieving this will require introductory training materials that explain how to use the tools as well as interactive interfaces that allow them to explore and understand the intermediate outputs that the system is relying on. For example, when the system calculates margins, the researcher needs to understand how it understands margins as compared to how a human might perceive them.

In summary, we found it is possible to extract visual features from document images that will support scholarly research questions and we have refined our understanding of the work that needs to be done. We see three main outstanding items: an engineering task to build more modular image analysis frameworks; a research task to move beyond physical pages to understand the semantic structure of different types of documents; and an infrastructure task to create the resources that will put this technology into the hands of the scholars best equipped to use it to address innovated research questions.

2.3 Data Analysis

Our initial data analysis goals included: confirming the relationship of the quantified measures of extracted features to the original document image; assessing which aspects of page design could be understood through these quantified measures; and developing approaches to

visualizing this information.

To confirm the relationship of the quantified measures to the original document image, a random sample of extracted features were matched to the original document images and assessed for general relevance and accuracy (at the level of the human eye, which perceives visual differences at a scale greater than the individual pixels analyzed by Tesseract). This assessment was very positive and suggested that there is potential for using this approach to large-scale visual understanding of printed documents.

As mentioned above, we selected poetry as our initial data set because the conventions of nineteenth-century printing visually distinguish on the page key features of poetic form, such as the grouping of lines in stanzas and the indenting of poetic lines to mark the rhyme scheme. Line indentation measures can be examined for regular variations that reflect these printing conventions for stanzaic form, as in Figure 2. Such measures can be used to determine the formal variety or regularity within a particular book. In future work, we plan to develop methods for using such measures to locate documents containing verse with particularly distinctive forms.

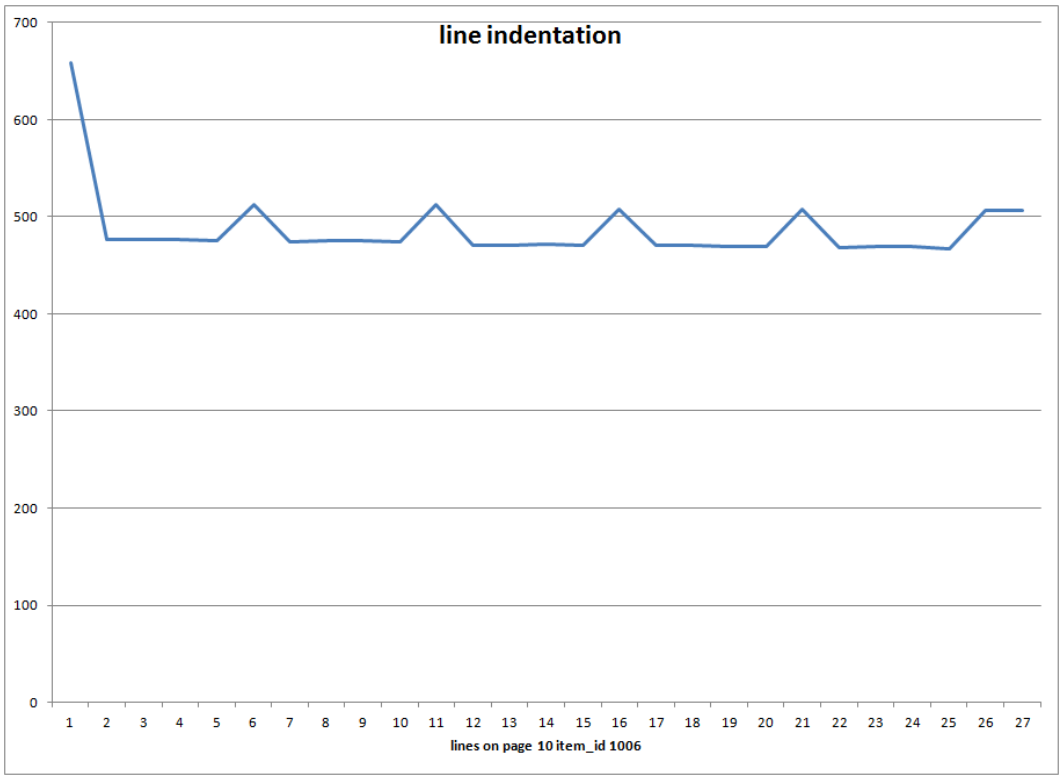
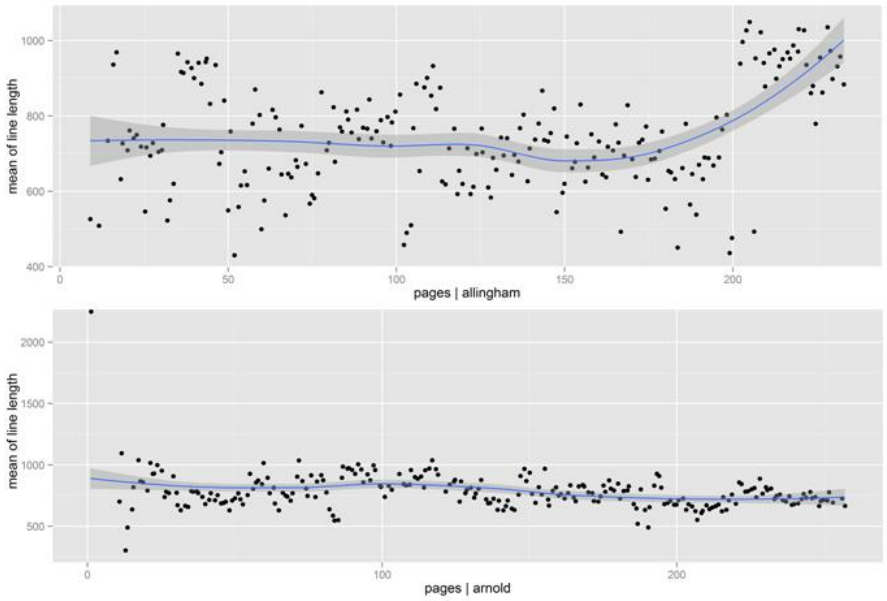


Figure 2

Because each poetic line is conventionally printed as a separate line of text on the page, line width is another feature that serves as an analytical link between the visual appearance of the page and the linguistic or literary features of the text. Measures of line width, for example, can be used to compare the relative consistency or variety in the forms of poetry included in two different books. Books with poetry all of the same or similar forms will have more consistent mean line width measurements per page than do books containing poetry of diverse forms. Such comparisons can be made purely at the quantitative level, for large-scale analytics, and can also be presented for human understanding through visualization, as shown in Figure 3.

This particular example also revealed some potential applications of our approach for research discovery. The unusual spike in line width values in the final pages of Allingham's book of poems reflected the publisher's advertisements included at the back of the book, which were printed in long prose blocks that filled the page. In future, we plan to develop analytical methods for using these extracted features to identify different kinds of texts in heterogenous document collections or within documents like nineteenth-century periodicals, which frequently printed poetry alongside prose and illustrations.



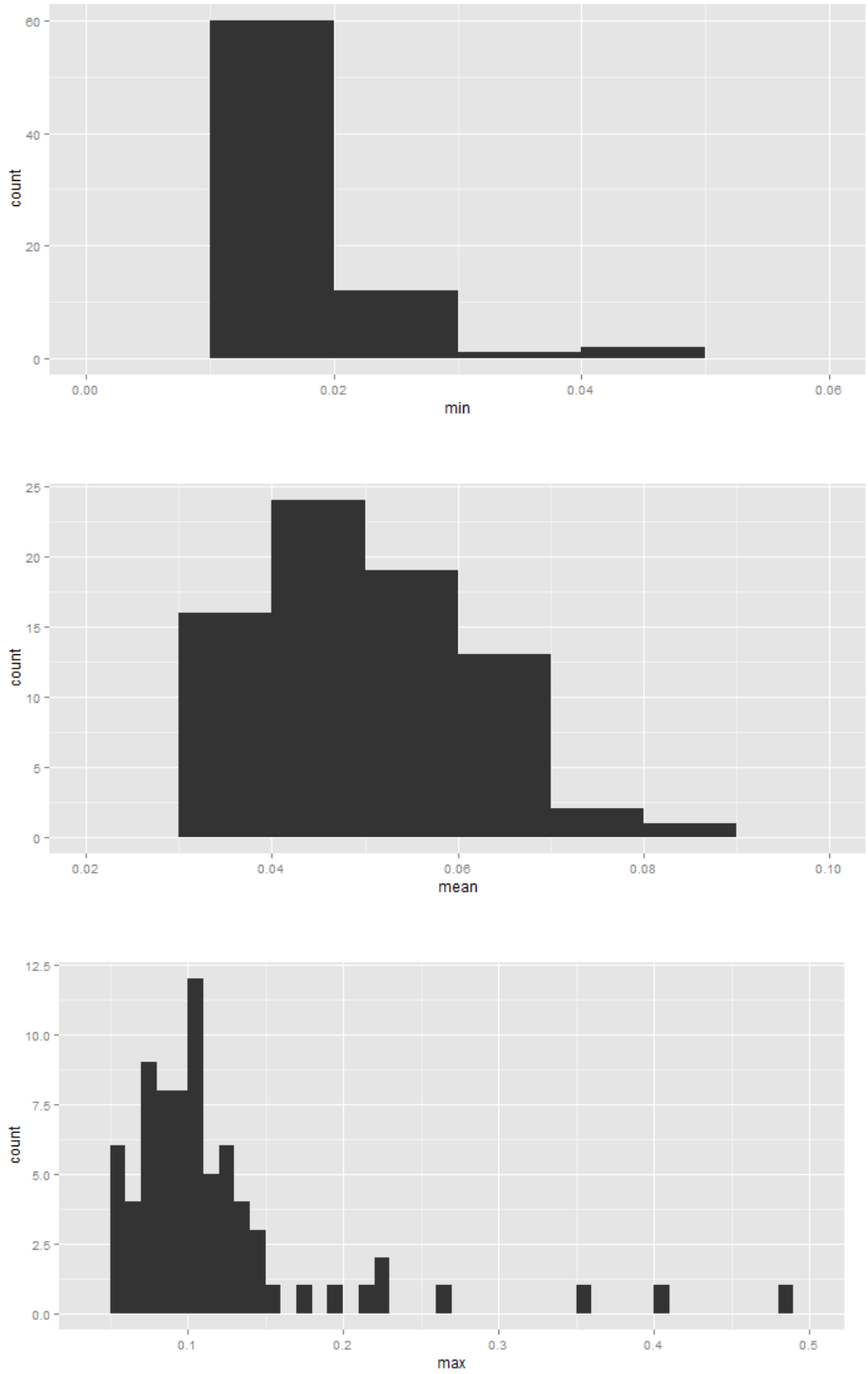
William Allingham, *Fifty Modern Poems* (Bell & Daldy, 1865)

Matthew Arnold, *New Poems* (Macmillan, 1867)

Comparing mean line widths per page in two volumes of poetry

Figure 3

Historians of nineteenth-century book design and typography frequently claim that printing was fairly standardized at mid-century due to economic constraints and technological efficiencies. Our initial data analysis demonstrated some strong consistencies in design among the small data set we were examining, but also suggests that much more work needs to be done to assess the best statistical models to use in examining data extracted from cultural documents. For instance, as Figure 4 indicates, the mean and minimum foreground ratio measures of the text density on the page show strong consistency in this data set. There is more variation in the measures of maximum foreground ratio, but much more statistical and historical work will need to be pursued to adequately interpret such results and assess their relevance.



Histograms of minimum, mean, and maximum foreground ratio values for set of 75 books

Figure 4

To examine the visual features of printed books through data visualizations such as these can also be understood as an act of critical deformance, an approach advocated by scholars like Stephen Ramsay and Jerome McGann that uses visualization techniques to help lead to new modes of humanist understanding.[4] We intend to pursue such strategic defamiliarizing of the printed book in future visualization work.

2.4 Presentation of Results

The project team jointly presented co-authored papers at the 2013 ADHO Digital Humanities conference and as part of the Big Humanities workshop held at the 2013 IEEE Big Data conference. In addition, aspects of this project were discussed in invited talks that Dr. Houston gave at Baylor University (2012), University of London (2013) and University of Kansas (2013), and in conference papers she presented at the North American Victorian Studies Association (2013) and the Modern Language Association (2014). Dr. Audenaert will also be presenting about the project at the April 2014 Texas Digital Humanities Consortium conference.

We were also invited to present the project in two graduate classes at Rice University: a Master Class in Digital History and an Introduction to Digital Humanities class.

3. Continuation of the Project

We plan to continue this research and have identified several directions for further work, including extending the document analysis to other types of printed documents and materials from other historical periods; further research into statistical models for large-scale understanding of printed documents; modifying and extending the system architecture to accommodate pluggable third-party components; research into the scalability of this approach to large-scale archives of cultural heritage materials; and research in scholarly user practices and the design of user interfaces.

To this end, we have recently been awarded a HathiTrust Research Center grant for a prototype project, as a subaward under the Workset Creation for Scholarly Analysis: Prototyping Project (WCSA) (http://worksets.htrc.illinois.edu/worksets/) funded by the Andrew W. Mellon Foundation.

---

[4] McGann, 105-136; Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism* (Urbana: U of Illinois P, 2011):32-57.

4. Grant Products

We developed a project website (visualpage.org) and will continue to post research results there. We posted some initial code for working with the open-source OCR engine Tesseract on Github (https://github.com/DART-Services/Visual-Page) and will continue to make code available as the project continues.